International Journal of Basic and Applied Sciences, 14 (2) (2025) 107-112

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/hgj17c14
Research paper

Text Mining and Natural Language Processing Frameworks for Enhanced Fake News Detection, Sentiment Analysis, and Automated Summarization in Social Media

K. Karnan [1] *, Dr. L. R. Aravind Babu [2]

[1] Research Scholar, Department of Computer Science and Information Science, Annamalai University, Tamil Nadu, India
[2] Department of Computer Science and Information Science, Annamalai University, Annamalai Nagar, Tamil Nadu, India
**Corresponding author E-mail: karnan.vdm@gmail.com*

Received: May 6, 2025, Accepted: May 18, 2025, Published: June 10, 2025

Abstract

Efficient text summarization, public sentiment analysis, and fake news detection have become difficult tasks due to the exponential growth of digital content. Sentiment analysis aids in assessing trends and public opinion, while fake news detection is crucial for combating false information. To alleviate information overload, automated text summarization extracts important information from long documents. This study examines three sophisticated Natural Language Processing (NLP) models: 1) The BiLSTM-based sentiment analysis model uses Word2Vec embeddings and bidirectional LSTM units to understand context better and classify text into positive, negative, or neutral sentiments. 2) Followed by a sigmoid classifier, to differentiate real from fake news, the BiLSTM-CNN-based fake news detection model combines a 1D CNN for spatial pattern recognition and BiLSTM for sequential feature extraction. 3) For extractive summarization, the hybrid extractive-abstractive summarization model uses TF-IDF-based sentence weighting for abstractive summarization it uses a Transformer-based encoder-decoder. The outcome is measured using metrics like BLEU and ROUGE. These models improve the online user experience , decision-making, and misinformation detection in text mining applications.

Keywords: *Fake News Detection; Sentiment Analysis; Text Summarization; Natural Language Processing (NLP); BiLSTM.*

1. Introduction

People's daily lives are impacted by fake news [1], which can change their opinions, feelings, and thoughts [2] and even lead them to make bad decisions. Because posting content on blogs, news websites, and social media is simple, people now have a harder time telling the difference between real and fake news [4]. Therefore, the issues of sentiment analysis, automated text summarization, and fake news detection are more important than ever. All these issues revolve around extracting meaningful information from huge amounts of unstructured text data, notwithstanding their differences. These issues are now being addressed, and how we analyze and manipulate digital content is being transformed through natural language processing (NLP) methods.

Fake news is a gross social issue due to its ability to mislead the public, influence elections, and spread negative stereotypes [5]. False data on social media can easily spread and reach a wide audience with minimal verification, and the implications can be fatal. Therefore, having a system that can reliably detect fake news is crucial. Manual verification techniques are time-consuming and impractical due to the volume of content generated online.

Automating fake news detection has shown promise with machine learning-based models, especially those that use deep learning techniques [6]. These models can accurately assess news stories' authenticity and credibility by utilizing textual features. The intricacy of language, the range of dishonest tactics employed in fake news, and the dynamic character of disinformation, however, make identifying fake news a difficult task.

Sentiment analysis has become a vital tool for comprehending trends, emotions, and public opinion at the same time. Because social media platforms have so much user-generated content, companies, government and other institutions are depending more and more on sentiment analysis to determine public sentiment and forecast future changes in society. In order to determine the emotions underlying user comments, reviews, or news articles, sentiment analysis divides text into different sentiment categories such as positive, negative, or neutral [7].

Organizations can make data-driven decisions, target audiences with content, and learn about customer feedback by comprehending sentiment. Sentiment analysis faces a major obstacle due to the complexity of human emotions, which calls for advanced models that can recognize linguistic subtleties and comprehend context. More reliable models that can manage such complexities must be developed because traditional sentiment analysis models frequently have trouble with irony, sarcasm, and context-dependent meanings.

In today's world of the internet, the problem of information overload is increasing in number. Numerous types of text content, including blog entries, research papers, social media status updates, and news articles, are uploaded to the internet daily. Because lengthy reports contain so much information, users might not be able to efficiently absorb the most important information. This problem is addressed by computerized text summarization, which reduces long texts into manageable logical chunks without sacrificing the most important information [8]. Extractive and abstractive summarization are the two main approaches. While abstractive summarizing creates new sentences that encapsulate the content of the source material, extractive summarizing involves choosing sentences or phrases from the source material. Due to their ability to combine the best features of both approaches, hybrid models are becoming more and more popular. They can effectively summarize text while maintaining coherence and contextuality.

Improvements in natural language processing (NLP), specifically the use of deep learning methods, have greatly enhanced the performance of sentiment analysis, text summarization, and detection of fake news in recent years. Using networks with Bidirectional Long Short-Term Memory (BiLSTM) is one approach that shows promise. Text classification and sentiment analysis are two tasks involving sequential data where BiLSTM models have shown great promise due to their ability to extract context from both past and future word sequences. BiLSTMs are a logical choice in sentiment analysis, where figuring out a sentence's sentiment usually requires understanding the relationships between words before and after a particular point in the text.

Convolutional neural networks (CNNs) are commonly employed in image processing, but they have also been applied to text classification tasks. CNNs aid in identifying spatial patterns in the input data when paired with BiLSTM models, which enhances the model's capacity to discern between authentic and fraudulent news. It has been shown that CNNs and BiLSTMs work especially well together for fake news detection, where the model must spot minute patterns and discrepancies in the text that might point to lies. These hybrid models offer a strong framework for detecting fake news by fusing the spatial pattern recognition capabilities of CNNs with the sequential feature extraction power of BiLSTMs.

Transformer-based architectures have also changed text summarization beyond these models. Transformer models such as the well-known BERT and GPT have demonstrated superior performance in a variety of NLP tasks, including summarization. These models are perfect for summarizing because they can understand long-range dependencies in text, which is important. After all, the main ideas of a lengthy text are often scattered throughout. Transformer-based models can be adjusted for both extractive and abstractive summarization, making them versatile instruments for creating concise perceptive summaries.

This paper investigates three advanced natural language processing models to address the issues of sentiment analysis, automated text summarization, and fake news detection. Text is categorized into three groups in the first model, a sentiment analysis system based on BiLSTM: neutral, negative, and positive. The second model is a framework for detecting fake news that is based on BiLSTM-CNN, which combines BiLSTM for sequential feature extraction with CNN for spatial pattern recognition. A Transformer-based encoder-decoder for abstractive summarization and TF-IDF-based extractive summarization are combined in the third model, which is an extractive-abstractive hybrid summarization system.

These models are at the forefront of NLP research and have the potential to significantly improve text mining applications by improving the detection of misinformation, providing useful data on public sentiment, and helping users navigate the overwhelming amount of information available in today's digital world. The goal of this study is to improve digital user experiences, decision-making, and the detection of false information by integrating these state-of-the-art NLP techniques.

2. Literature review

Disseminated information and the process of its dissemination create a significant challenge in quickly detecting this content, thereby emphasizing the necessity of automatically recognizing false news. To address this, [9] introduced an automatic fake news identification method for the Chrome environment, making the detection of fake news on Facebook feasible. Specifically, this uses a number of Facebook account-related features in addition to several news content attributes to analyze the accounts' attributes using deep learning.

[10] Presented a CNN + bidirectional LSTM ensemble network to generate diverse information for differentiating between fake and real news when comparing numerous state-of-the-art techniques, and to obtain new samples such as PolitiFact. Among the many cutting-edge techniques are ensemble approaches, convolutional neural networks (CNNs), long short-term memories (LSTMs), and attention mechanisms. This study creates several datasets for real and fake news articles by collecting 1356 news examples from various users via Twitter and media websites like PolitiFact. According to the research findings, the CNN + bidirectional LSTM ensemble network with an attention mechanism achieved the best accuracy, ranging from 88 to 78%. Though not particularly encouraging, the results were satisfactory.

[11] Implemented a multimodal strategy utilizing Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to categorize fake news articles, which resulted in remarkable performance. We focused on a dataset containing 12 distinct categories of news articles and employed linguistic cue techniques alongside machine learning. We categorized news based on its origin and its records (like domain name and/or author name) with a bimodal CNN and LSTM. The model, through credible news sources, classifies trustworthy news articles with an accuracy of 99.7% on the training set and 97.5% on the test set. Nevertheless, since fake news can still be released on a credible domain, we had to consider additional factors such as news headlines.

In comparison to Random Forest (RF) and Passive Aggressive classifiers with and without text pre-processing, [12] computed performance accuracy for the BuzzFeed News dataset using the suggested Naïve Bayes (NB) classifier. When applied to the body feature matrix without pre-processing, Naive Bayes proved to be the most successful model in predicting real news from fraudulent news with 99% accuracy. In this investigation, the Naive Bayes classifier performed better than the Random Forest and Passive Aggressive classifiers when combined with Gradient Boost.

[13] Suggested combining RF and SVM classification methods to increase F1-score, recall, accuracy, and precision. Additionally, the authors analyze the performance of the hybrid RFSVM and present a comparative study of several news categories using different machine-learning models. In the realms of accuracy, precision, recall, and F1-score, hybrid RFSVM surpassed others such as Random Forest (RF), naïve Bayes (NB), SVMs, and XGBoost by 8% to 16%.

[14] Discussed the application of Transformer-based models, i.e., BART and T5, for abstractive podcast transcript summarization. Fine-tuning the models on a massive dataset of podcast episodes, the research compares their performance on machine metrics and human assessment. The findings indicate that BART is superior on ROUGE-1 and ROUGE-L scores, but T5 is superior on semantic meaning preservation, indicating the relative strengths of the two models in different aspects of summarization.

[15] Presents a new fake news detection method using textual, visual, and social contextual information through Graph Neural Networks (GNNs) with the aid of attention mechanisms. The proposed Multimodal Fake News Detector (MFND) utilizes the intricate interdepend-

encies among various modalities of data to improve detection accuracy. In benchmarking using FakeNewsNet and Sina Weibo, the model performs better than current methods, thereby highlighting the efficacy of multimodal integration in the fight against misinformation.
There have been various studies that have been successful in detecting misinformation, sentiment analysis, and summarizing using different models. There remains scope for enhancement in these tasks by integrating their strengths in a single framework. For further expansion of the previous studies, the next section discusses our approach using advanced natural language processing models such as BiLSTM, CNN, and Transformers to analyze and summarize social media content more effectively.

3. Methodology

The proposed approach enhances sentiment analysis, fake news detection, and automated summarization by integrating Natural Language Processing (NLP) and deep learning models. Convolutional Neural Networks (CNN), Transformer-based models and Bidirectional Long Short-Term Memory (BiLSTM) are combined in this method to improve text classification and summarization performance. Modern word embedding methods, such as Word2Vec and Term Frequency-Inverse Document Frequency (TF-IDF) are used to train the models on preprocessed data to extract meaningful patterns from text data.
Mathematically, given an input sequence $X = (x_1, x_2, \ldots, x_n)$, where x_i represents tokenized words, the embeddings $E(x_i)$ are generated using Word2Vec. The BiLSTM layer processes the sequence bidirectionally:

$$h_t^f = \mathrm{LSTM}_f\left(E(x_t), h_{t-1}^f\right), h_t^b = \mathrm{LSTM}_b\left(E(x_t), h_{t+1}^b\right) \quad (1)$$

Where h_t^f and h_t^b signify forward and backward, hidden conditions, respectively. These are concatenated to get the final representation:

$$h_t = \left[h_t^f, h_t^b\right] \quad (2)$$

For sentiment classification, a softmax activation function is used:

$$P(y|X) = \mathrm{softmax}(Wh_t + b) \quad (3)$$

For fake news detection, a CNN layer extracts spatial features using a convolutional filter W_c:

$$f_t = \mathrm{ReLU}(W_C * h_t + b_c) \quad (4)$$

Where $*$ denotes the convolution operation, followed by max pooling to downsample features before passing them to a sigmoid classifier. The hybrid extractive-abstractive summarization model first applies TF-IDF weighting to extract key sentences. Given a term t in a document d, its TF-IDF score is computed as:

$$TF - IDF(t, d) = TF(t, d) \times \log\left(\frac{N}{DF(t)}\right) \quad (5)$$

Where N is the total number of documents and $DF(t)$ is the number of documents containing t. The selected sentences are then passed to a Transformer-based encoder-decoder model, which generates an abstractive summary. The final summary is evaluated using ROUGE and BLEU metrics, ensuring coherence and informativeness.
In this case, N represents the total number of documents, and DF(t) represents the number of documents that contain t. An abstractive summary is then produced by a Transformer-based encoder-decoder model using the chosen sentences. ROUGE and BLEU measures are used to assess the final summary, guaranteeing its coherence and informativeness.
Figure 1 shows the BiLSTM sentiment analysis architecture. The BiLSTM-based sentiment analysis model uses Long Short-Term Memory (LSTM) networks with bidirectional processing ability to categorize text data into positive, negative, or neutral sentiments.

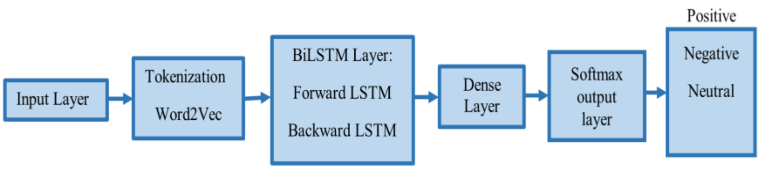

Fig. 1: BiLSTM-Based Sentiment Analysis Architecture.

The input text is first preprocessed, which includes tokenization in which Word2Vec embeddings are used to translate words into numerical representations. The model can comprehend contextual meaning because these embeddings capture the semantic relationships between words.
A BiLSTM layer comprising both forward and backward LSTM units receives the tokenized input after that. To ensure that the model captures dependencies from both directions and enhances contextual understanding, the forward LSTM processes the sequence from the past to the future while the backward LSTM processes it in reverse.
A dense layer then refines the feature representations that the recurrent units extracted from the BiLSTM output. Lastly, a softmax layer is used to categorize the sentiment into three groups: neutral, negative, and positive.
By utilizing the power of bidirectional sequence modelling, this architecture improves sentiment analysis performance, making it appropriate for uses like opinion mining, social media analysis, and customer feedback assessment.
The BiLSTM with CNN hybrid model for detecting fake news is shown in Figure 2. The purpose of the BiLSTM-CNN-based fake news detection model is to efficiently extract spatial and sequential features from textual data for binary classification.

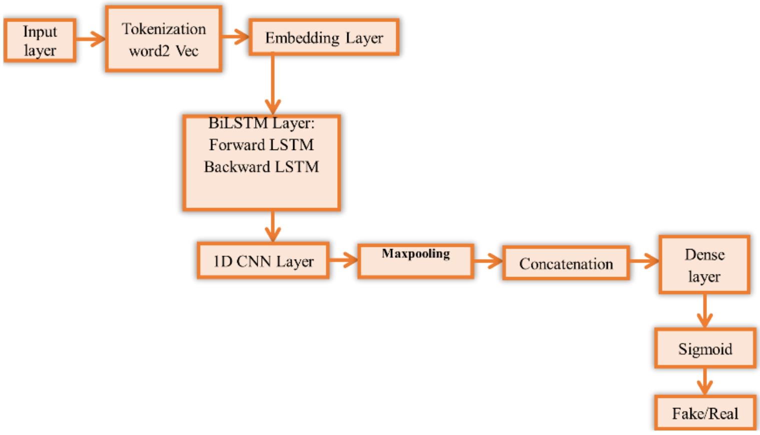

Fig. 2: BiLSTM-CNN-Based Fake News Detection Architecture.

Tokenization is used to preprocess raw text data in the input layer, and Word2Vec embeddings, which map words to dense vector spaces while maintaining semantic relationships, are used to transform the data into numerical representations.
After that, these embeddings are sent to a Bidirectional Long Short-Term Memory (BiLSTM) layer, which reads the sequence both forward and backward to help the model understand contextual meaning and long-range dependencies.
A 1D Convolutional Neural Network (CNN) is used to extract spatial features and local relationships in the text after the BiLSTM layer. The next step is a max pooling layer, which reduces dimensionality while maintaining the most important characteristics. An end-to-end feature representation that combines sequential and spatial features is created by concatenating the outputs from the BiLSTM and max pooling layer. To classify the input as either fake or real news, this combined representation passes through a dense layer and then the output layer's sigmoid activation function.
Figure 3 illustrates the hybrid extractive-abstractive text summarization model, which combines extractive and abstractive summarization techniques to produce high-quality summaries.

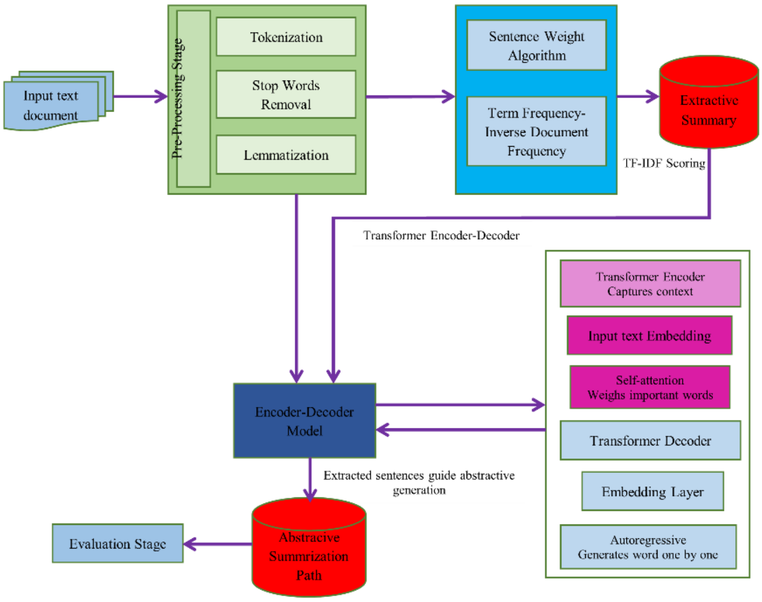

Fig. 3: Hybrid Extractive-Abstractive Summarization Architecture.

Preprocessing is the first step in the process, during which the input text document is subjected to necessary text-cleaning procedures like lemmatization, tokenization, and stop-word removal. The extractive summary database is created by running the preprocessed text through a sentence weighting algorithm that uses Term Frequency-Inverse Document Frequency (TF-IDF) to find and extract the most pertinent sentences.
For abstractive summarization, the preprocessed text is simultaneously fed straight into an encoder-decoder model based on the Transformer.
Using text embeddings and self-attention, the encoder in the encoder-decoder model processes the input and extracts rich contextual relationships from the document. An autoregressive mechanism and an embedding layer are used by the decoder to generate a logical, human-like summary. By providing essential context, the extractive summary also aids in refining the abstractive summary. Lastly, the evaluation step uses metrics like ROUGE and BLEU to evaluate the quality of the generated summary.

4. Results and discussion

This section provides the experimental results from the three NLP models: the BiLSTM-based sentiment analysis model, the hybrid extractive-abstractive summarization model, and the BiLSTM-CNN-based fake news detection model. Each model's performance was estimated using pertinent metrics, and the outcomes were compared with previous methods found in the literature.

The proposed model's performance is evaluated experimentally with benchmark datasets for text summarization, sentiment analysis, and fake news detection. It is evaluated using common evaluation metrics: accuracy, precision, recall, and F1-score for sentiment analysis and the identification of fake news; ROUGE-1, ROUGE-2, ROUGE-L, and BLEU for text summarization. The used datasets are here: For Sentiment Analysis, Sentiment140, IMDB Reviews are used, for Fake News Detection: FakeNewsNet, LIAR dataset is used, for Text Summarization: DailyMail dataset, CNN is used.
Tokenization, stop-word removal, and lemmatization were used as preprocessing techniques for each dataset. 20% of the dataset was used for testing, and the remaining 80% was used for training. TensorFlow and PyTorch were used to implement the models, and the Adam optimizer was used for optimization with a learning rate of 0.001.

4.1. Sentiment analysis

The BiLSTM-based sentiment analysis model's performance was measured using accuracy, precision, recall, and F1-score. The model was evaluated on a labeled dataset comprising a variety of sentiment expressions, including sarcasm and sophisticated language structures. The results are summarized in Table 1.

Table 1: Performance of BiLSTM Sentiment Analysis Model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LSTM	83.2	81.9	82.5	82.2
CNN	85.4	83.7	84.5	84.1
BiLSTM	89.3	88.1	88.7	88.4

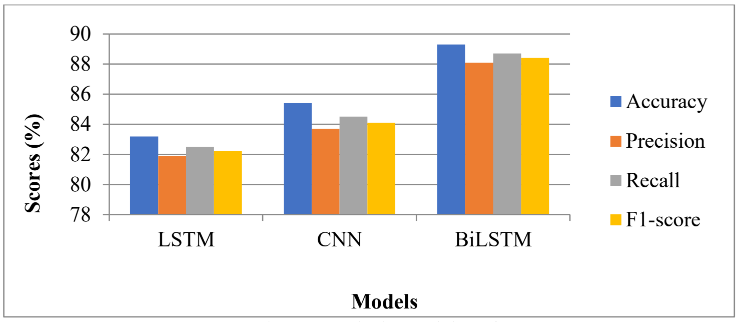

Fig. 4: Sentiment Analysis Model Performance.

Figure 4 represents the model performance of sentiment analysis. The outcomes show that the BiLSTM model outperforms conventional machine learning models like LSTM and CNN in terms of capturing context and sentiment dependencies.

4.2. Fake news detection

Accuracy, precision, recall, and F1-score were used to assess the BiLSTM-CNN-based fake news detection model. Benchmark datasets like LIAR and FakeNewsNet were used to train and evaluate the model. Table 2 presents the effectiveness of the proposed approach.

Table 2: Performance of BiLSTM-CNN Fake News Detection Model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Random Forest	74.8	72.6	73.4	73.0
Naïve Bayes	78.1	76.5	77.2	76.8
CNN	85.2	84.3	84.8	84.5
BiLSTM-CNN	91.6	90.2	90.9	90.5

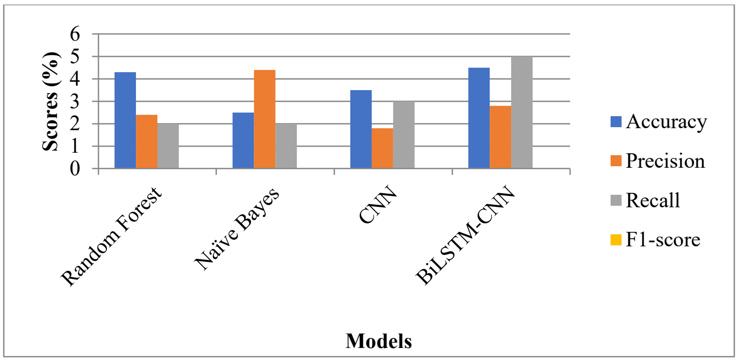

Fig. 5: Fake News Detection Model Performance.

Figure 5 depicts the fake news detection model performance. From this result, it is clear that the proposed BiLSTM-CNN model outperforms the compared models.

4.3. Text summarization

ROUGE and BLEU measures were used to assess the hybrid extractive-abstractive summarization model. The BLEU assesses fluency and coherence while the ROUGE scores quantify the content overlap between generated and reference summaries. The findings are shown in Table 3.

Table 3: Performance of Hybrid Text Summarization Model

Model	ROUGE-1 (%)	ROUGE-2 (%)	ROUGE-L (%)	BLEU (%)
Extractive (TF-IDF)	45.3	33.2	42.5	38.1
Abstractive (Transformer)	51.7	37.8	47.2	41.5
Hybrid	58.9	44.1	53.4	47.3

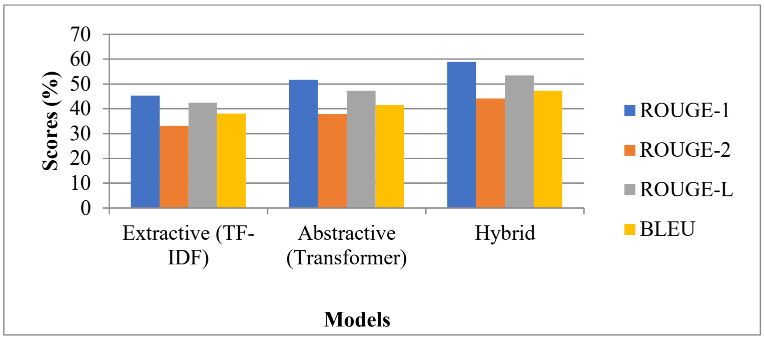

Fig. 6: Text Summarization Model Performance.

Figure 6 represents the text summarization of the proposed model. The deep learning models that have been suggested improve performance on all NLP tasks, with an accuracy rate of 89.3% the BiLSTM-based sentiment analysis model outperforms CNN and LSTM. The BiLSTM-CNN fake news detection model is very good at detecting misinformation with an accuracy of 91.6%. In order to produce coherent summaries, the hybrid extractive-abstractive text summarization model achieves a BLEU score of 47.3% and a ROUGE-1 score of 58.9%. In general, hybrid deep learning models enhance text summarization, sentiment analysis, and fake news detection accuracy and robustness.

5. Conclusion

Experimental results show that the suggested NLP models are effective in text summarization, sentiment analysis, and fake news detection. The BiLSTM-based sentiment analysis model outperformed CNN-based and LSTM-based models in terms of accuracy (89.3%), demonstrating its capacity to identify contextual dependencies for sentiment classification. Similarly, the BiLSTM-CNN fake news detection model demonstrated the benefit of combining sequential and spatial feature extraction by outperforming conventional classifiers with an accuracy of 91.6%. High ROUGE and BLEU scores were attained by the hybrid extractive-abstractive summarization model, guaranteeing coherence and relevance in summary creation. These findings indicate the feasibility of more robust text mining tools in web-based environments, but actual deployment will likely continue to encounter issues with the complexity of computation, scalability for big data, and flexibility to support multiple languages. Future research can concentrate on maximizing model performance in low-resource and real-time settings. The framework's practicality will be improved by adding support for multilingual and multimodal data.

References

[1] Shrivastava, G., Kumar, P., Ojha, R. P., Srivastava, P. K., Mohan, S., & Srivastava, G. (2020). Defensive modeling of fake news through online social networks. IEEE Transactions on Computational Social Systems, 7(5), 1159-1167. https://doi.org/10.1109/TCSS.2020.3014135.

[2] Xu, K., Wang, F., Wang, H., & Yang, B. (2019). Detecting fake news over online social media via domain reputations and content understanding. Tsinghua Science and Technology, 25(1), 20-27. https://doi.org/10.26599/TST.2018.9010139.

[3] Habib, A., Asghar, M. Z., Khan, A., Habib, A., & Khan, A. (2019). False information detection in online content and its role in decision making: a systematic literature review. Social Network Analysis and Mining, 9, 1-20. https://doi.org/10.1007/s13278-019-0595-5.

[4] Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining, 13(1), 30. https://doi.org/10.1007/s13278-023-01028-5.

[5] Feng, Z. (2019). Hot news mining and public opinion guidance analysis based on sentiment computing in network social media. Personal and Ubiquitous Computing, 23(3), 373-381. https://doi.org/10.1007/s00779-018-01192-y.

[6] Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. Sensors, 20(18), 5212. https://doi.org/10.3390/s20185212.

[7] Yu, J., & Qi, C. (2025). Machine Learning-Based Sentiment Analysis in English Literature: Using Deep Learning Models to Analyze Emotional and Thematic Content in Texts. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3553386.

[8] Liu, Z. (2025). From Extractive to Generative: An Analysis of Automatic Text Summarization Techniques. In ITM Web of Conferences (Vol. 73, p. 02008). EDP Sciences. https://doi.org/10.1051/itmconf/20257302008.

[9] Sahoo SR, Gupta BB (2021) Multiple features based approach for automatic fake news detection on social networks using deep learning. Appl Soft Comput 100(106983):106983. https://doi.org/10.1016/j.asoc.2020.106983.

[10] Hannah Nithya S, Sahayadhas A (2022) Automated fake news detection by LSTM enabled with optimal feature selection. J Inf Knowl Manag 21(03). https://doi.org/10.1142/S0219649222500368.

[11] Javed Awan M, Shehzad F, Muhammad H, Ashraf M (2020) Fake news classification bimodal using convolutional neural network and long short-term memory. Int J Emerg Technol 11(5):197–204

[12] Matemilola, A. S., & Aliyu, S. (2024). Development of an enhanced naive bayes algorithm for fake news classification. Science World Journal, 19(2), 512-517. https://doi.org/10.4314/swj.v19i2.28.

[13] Dev, D. G., & Bhatnagar, V. (2024). Hybrid RFSVM: Hybridization of SVM and Random Forest Models for Detection of Fake News. Algorithms, 17(10), 459. https://doi.org/10.3390/a17100459.

[14] Saxena, P., & El-Haj, M. (2023, September). Exploring abstractive text summarisation for podcasts: a comparative study of BART and T5 models. In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (pp. 1023-1033). https://doi.org/10.26615/978-954-452-092-2_110.

[15] Li, Z. (2025). Multimodal fake news detection using graph neural networks and attention mechanisms. Advances in Engineering Innovation, 15, 63-73. https://doi.org/10.54254/2977-3903/2025.20827.